Excel Spreadsheet Data Entry Tips

Prepared by the ***Center for Biostatistics****, Department of Biomedical Informatics*
The Ohio State University

Revised 04/15/2015

Note:

The following document presents some general tips and guidelines for research data entered into Excel. These tips are designed to avoid common problems and minimize the time spent 'cleaning' the data before actual analysis.

While these rules cover some basic aspects of data entry, we strongly recommend consulting with a statistician before starting to collect your data.

Excel is currently a popular choice for data entry/management for small research projects. However, Excel may not necessarily be the best/most efficient way of recording your data; a statistician can recommend other choices which may be better suited to your particular needs.

This document was prepared by the **Center for Biostatistics**. Please contact us if you have questions/comments about this guide or would like to speak with a statistician regarding your particular research project.

Center for Biostatistics:
Website: http://www.biostatistics.osu.edu
Email: biostatistics@osumc.edu
Phone: (614) 293-6899

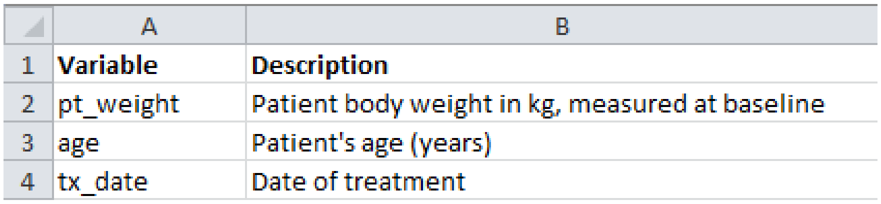
Excel Spreadsheet Data Entry Tips

Prepared by the Center for Biostatistics, Department of Biomedical Informatics
The Ohio State University

Revised: 04/15/2015

Excel Spreadsheet Data Entry Tips:

1. **Variable Names:**
   - Enter variable names in the first row of the spreadsheet.
   - Do not put spaces in the name. Use the underscore "_" character instead (e.g., "`body_weight`" instead of "`body weight`").
   - Keep variable names simple and short (e.g., '`pt_weight`' or '`bodywt`' instead of '`patient_body_weight_in_kg_measured_at_baseline`')
   - You can create a key to put more information about each variable on a separate sheet if desired:

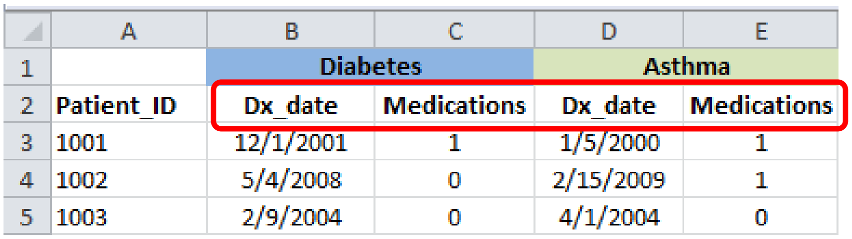


	A	B
1	**Variable**	**Description**
2	pt_weight	Patient body weight in kg, measured at baseline
3	age	Patient's age (years)
4	tx_date	Date of treatment

Be sure to put the units of measurement (e.g., kg, years) in the description where applicable!

   - The first character of the variable name should be a letter, not a number (e.g., '`week1`' instead of '`1st_week`').
   - No special characters (e.g., !, @, *) in the name.
   - Make sure each variable name is unique. Do not use merged cells to differentiate variables.

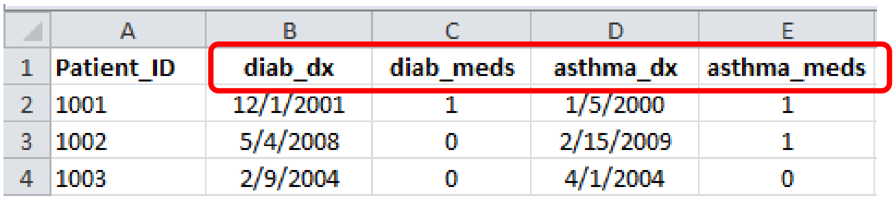
So instead of this…



	A	B	C	D	E
1		**Diabetes**		**Asthma**	
2	**Patient_ID**	**Dx_date**	**Medications**	**Dx_date**	**Medications**
3	1001	12/1/2001	1	1/5/2000	1
4	1002	5/4/2008	0	2/15/2009	1
5	1003	2/9/2004	0	4/1/2004	0

The variable names '`dx_date`' and '`Medications`' are repeated in the 2nd row.

Arrange your data like this…



	A	B	C	D	E
1	**Patient_ID**	**diab_dx**	**diab_meds**	**asthma_dx**	**asthma_meds**
2	1001	12/1/2001	1	1/5/2000	1
3	1002	5/4/2008	0	2/15/2009	1
4	1003	2/9/2004	0	4/1/2004	0

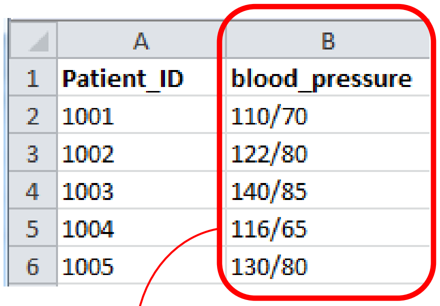
2. ID Variable:

- Always include an ID variable on each sheet of your workbook so that variables are properly associated with each subject.
  - Note that the ID needs to be unique for each subject in your study!

Do not use MRN, patient names, social security numbers, or any other identifying information that would violate HIPAA rules as ID variables – create your own study ID instead and be sure to keep a key so that you can match the study ID with the original identifying information.

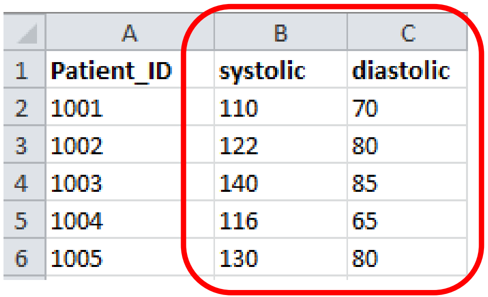
3. General Data Entry Rules:

- One row per subject (please see '8. Multiple Observations Per Subject' for exceptions).
- One column per variable.
- One value per cell (please see '5. Multiple Responses' for more information).
  - Special case: values composed of multiple components, such as **blood pressure**



	A	B
1	**Patient_ID**	**blood_pressure**
2	1001	110/70
3	1002	122/80
4	1003	140/85
5	1004	116/65
6	1005	130/80

NO



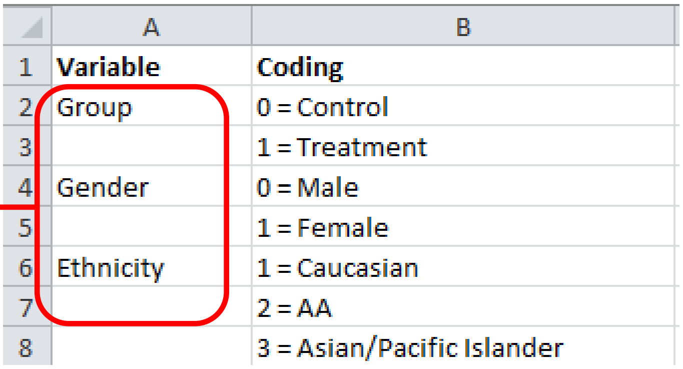
	A	B	C
1	**Patient_ID**	**systolic**	**diastolic**
2	1001	110	70
3	1002	122	80
4	1003	140	85
5	1004	116	65
6	1005	130	80

YES

Instead of putting both systolic and diastolic pressures in one column, create separate columns for each component.

- Avoid text for values if possible – use numbers instead (e.g., 0 for Male and 1 for Female).
- If you use text values, be careful about spelling/capitalization!!
  - In our statistical programs, 'Male' is not the same as 'male' or 'M' – use only one form in your data entry and be consistent!

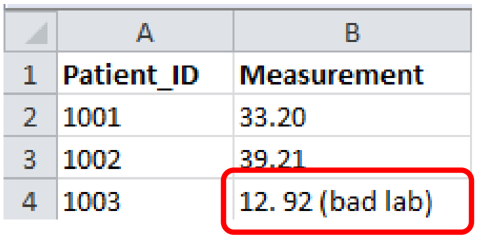
- If you do use numbers to represent text, we recommend creating a key on a separate sheet:



	A	B
1	**Variable**	**Coding**
2	Group	0 = Control
3		1 = Treatment
4	Gender	0 = Male
5		1 = Female
6	Ethnicity	1 = Caucasian
7		2 = AA
8		3 = Asian/Pacific Islander

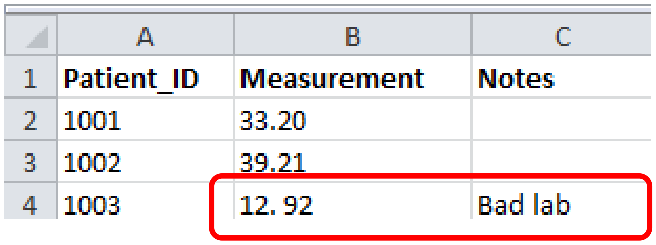
Be sure that the variable names in your key match the names in the spreadsheet!

- Any extra text or notes should go in a separate column, not within the variables themselves.



	A	B
1	**Patient_ID**	**Measurement**
2	1001	33.20
3	1002	39.21
4	1003	12. 92 (bad lab)

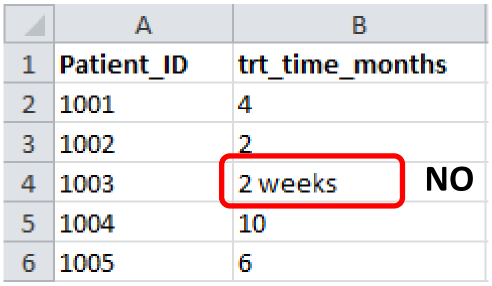
NO



	A	B	C
1	**Patient_ID**	**Measurement**	**Notes**
2	1001	33.20	
3	1002	39.21	
4	1003	12. 92	Bad lab

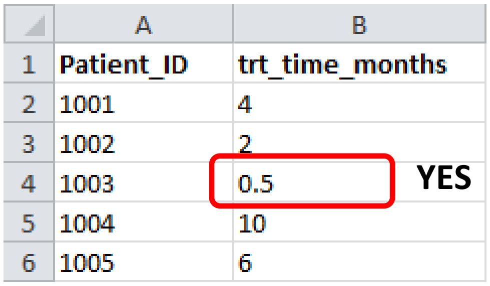
YES

- For numeric variables, please be consistent with the units:



	A	B
1	**Patient_ID**	**trt_time_months**
2	1001	4
3	1002	2
4	1003	2 weeks
5	1004	10
6	1005	6

NO



	A	B
1	**Patient_ID**	**trt_time_months**
2	1001	4
3	1002	2
4	1003	0.5
5	1004	10
6	1005	6

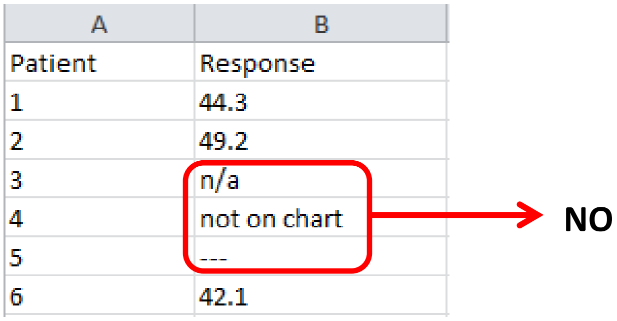
YES

In this example, the unit of measurement is 'months'. Therefore, we want to make sure that all entries are measured in months (not weeks, days, or any other units of time).

- **Do not** use the following to organize your data:
  - Color coding
  - Merged cells
  - Blank rows/columns

4. Missing Data

- Leave blank or code with an identifier that does not match any other numerical value entered (e.g., -9999).
- Do not use text to represent missing data, especially if your variable is numeric:

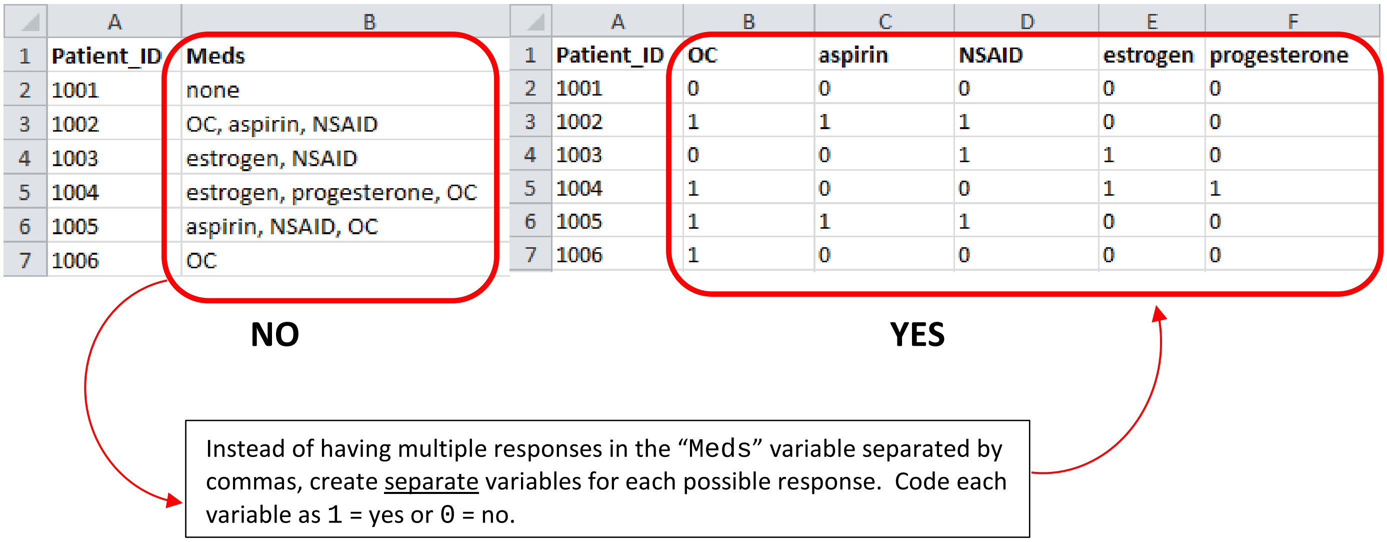


A	B
Patient	Response
1	44.3
2	49.2
3	n/a
4	not on chart
5	---
6	42.1

NO

5. Multiple Responses

- When a question/variable has multiple responses that are not mutually exclusive, we recommend you create separate variables for each response.
- For example, suppose we have a variable “Meds” listing all of the medications a patient was taking:



	A	B
1	**Patient_ID**	**Meds**
2	1001	none
3	1002	OC, aspirin, NSAID
4	1003	estrogen, NSAID
5	1004	estrogen, progesterone, OC
6	1005	aspirin, NSAID, OC
7	1006	OC

NO

	A	B	C	D	E	F
1	**Patient_ID**	**OC**	**aspirin**	**NSAID**	**estrogen**	**progesterone**
2	1001	0	0	0	0	0
3	1002	1	1	1	0	0
4	1003	0	0	1	1	0
5	1004	1	0	0	1	1
6	1005	1	1	1	0	0
7	1006	1	0	0	0	0

YES

Instead of having multiple responses in the “Meds” variable separated by commas, create separate variables for each possible response. Code each variable as 1 = yes or 0 = no.

6. Dates

- Please use MM/DD/YYYY format (e.g., 12/12/2014).
- Be consistent when entering dates, particularly with the 4 digit year! (e.g., do not put both 12/12/2014 and 12/12/14)

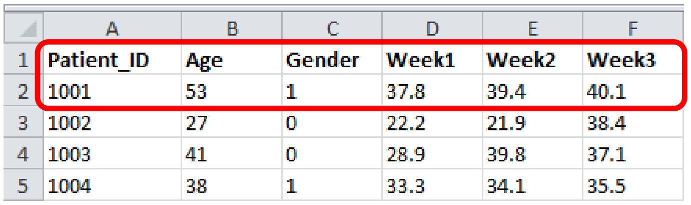
7. Calculated Data

- We always prefer the original (‘raw’) data over calculated totals, formulas, ‘normalized’ data, etc.
    - For example, please provide actual dates instead of calculated days between measurements, raw CT values rather than fold changes.
- We can calculate these quantities easily in our statistical programs.

8. Multiple Observations Per Subject

- When subjects have multiple observations (e.g., time points, replicates, etc.), we generally prefer that the data are arranged so that there are multiple rows per subject, one row for each observation.
- Be sure to repeat the ID value as well as any other variables that are associated with the subject that remain constant (e.g., race, gender).

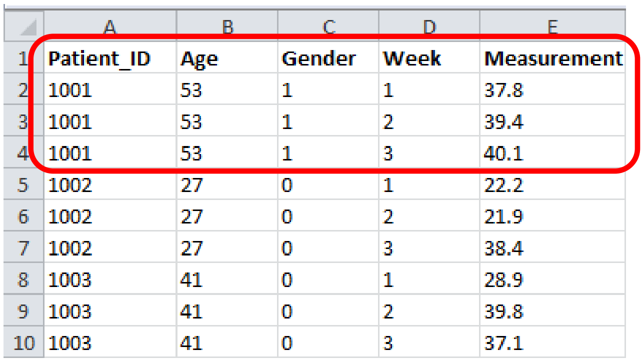
So instead of this…



	A	B	C	D	E	F
1	**Patient_ID**	**Age**	**Gender**	**Week1**	**Week2**	**Week3**
2	1001	53	1	37.8	39.4	40.1
3	1002	27	0	22.2	21.9	38.4
4	1003	41	0	28.9	39.8	37.1
5	1004	38	1	33.3	34.1	35.5

One row per patient, each measurement in a separate column (‘`Week1`’ – ‘`Week3`’)

Arrange your data like this…



	A	B	C	D	E
1	**Patient_ID**	**Age**	**Gender**	**Week**	**Measurement**
2	1001	53	1	1	37.8
3	1001	53	1	2	39.4
4	1001	53	1	3	40.1
5	1002	27	0	1	22.2
6	1002	27	0	2	21.9
7	1002	27	0	3	38.4
8	1003	41	0	1	28.9
9	1003	41	0	2	39.8
10	1003	41	0	3	37.1

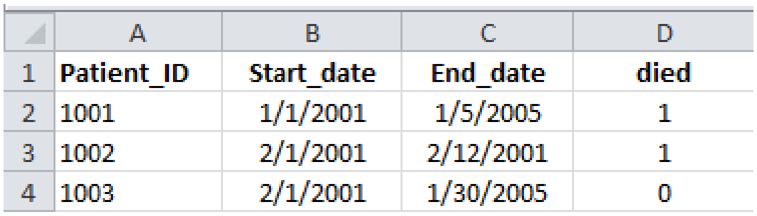
One row for each observation. All measurements are in one column (‘`Measurement`’), with another column (‘`Week`’) to identify the week. Note that the ID, age, and gender variables need to be filled in for each observation.

Please talk with your statistician about your particular situation before entering data!

9. Survival Data

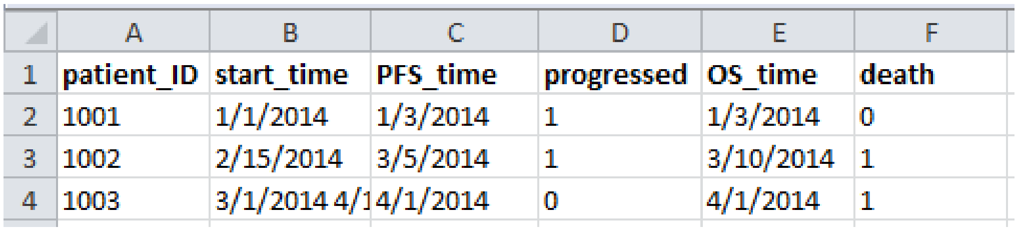
- **Do not** provide summary data by time point!

- For each subject, provide the following information:
  - Start date/time
  - End date/time – this corresponds to either the date/time the subject had the event of interest, or the last date/time of the study.
  - Status at end date/time:
    - 1 = died/had the event of interest
    - 0 = censored/did not have the event of interest



	A	B	C	D
1	**Patient_ID**	**Start_date**	**End_date**	**died**
2	1001	1/1/2001	1/5/2005	1
3	1002	2/1/2001	2/12/2001	1
4	1003	2/1/2001	1/30/2005	0

- If your survival data are more complicated (e.g., you want to look at overall survival as well progression-free or disease-free survival, or want to consider competing risks), be sure to talk with your statistician about the best way to record the information:
  - Dates are always preferred over calculated times.
  - When possible, each event/time should be put in separate columns.



	A	B	C	D	E	F
1	**patient_ID**	**start_time**	**PFS_time**	**progressed**	**OS_time**	**death**
2	1001	1/1/2014	1/3/2014	1	1/3/2014	0
3	1002	2/15/2014	3/5/2014	1	3/10/2014	1
4	1003	3/1/2014 4/1	4/1/2014	0	4/1/2014	1

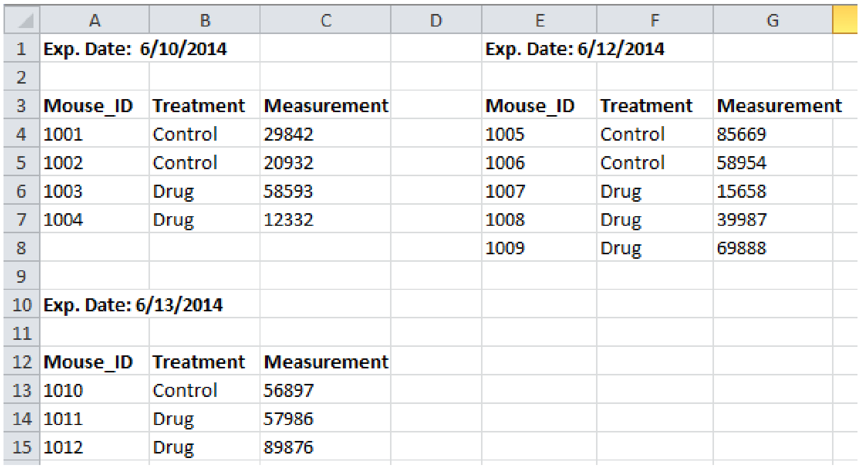
In this example, both overall survival (OS) and progression free survival (PFS) are of interest. Note that we have separate columns – time and an indicator – for each event.

10. Multiple Datasets

- Only include one dataset per sheet. Do not put unrelated sets of data on the same page.

- **<u>Exception</u>:** If you are collecting the <u>same information</u> in various datasets (e.g., running the same experiment over different time points/batches, collecting the same information in different treatment groups), you can arrange the data on one sheet. In this case, please do not put the data in 'blocks':

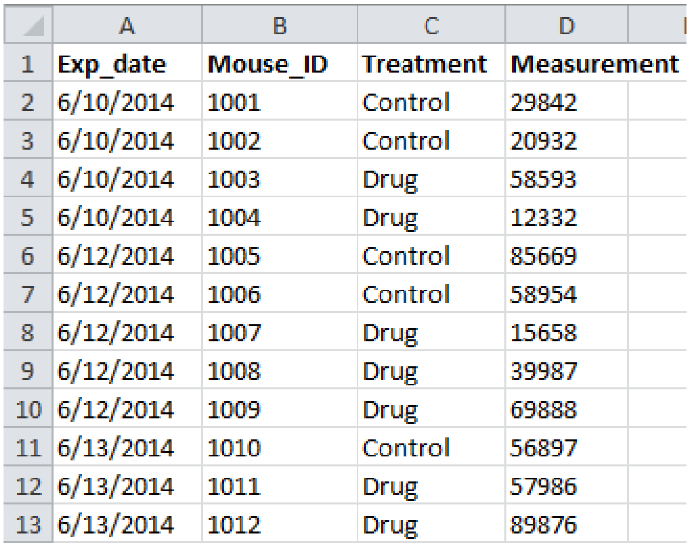
So instead of this…



	A	B	C	D	E	F	G
1	**Exp. Date: 6/10/2014**				**Exp. Date: 6/12/2014**		
2							
3	**Mouse_ID**	**Treatment**	**Measurement**		**Mouse_ID**	**Treatment**	**Measurement**
4	1001	Control	29842		1005	Control	85669
5	1002	Control	20932		1006	Control	58954
6	1003	Drug	58593		1007	Drug	15658
7	1004	Drug	12332		1008	Drug	39987
8					1009	Drug	69888
9							
10	**Exp. Date: 6/13/2014**						
11							
12	**Mouse_ID**	**Treatment**	**Measurement**				
13	1010	Control	56897				
14	1011	Drug	57986				
15	1012	Drug	89876				

In this example, the same experiment was run in batches on three different dates. Note how the data are grouped in 'blocks' according to the date of the experiment. Therefore, values for '`Mouse_ID`', '`Treatment`', and '`Measurement`' are contained in more than one column.

Arrange your data like this…

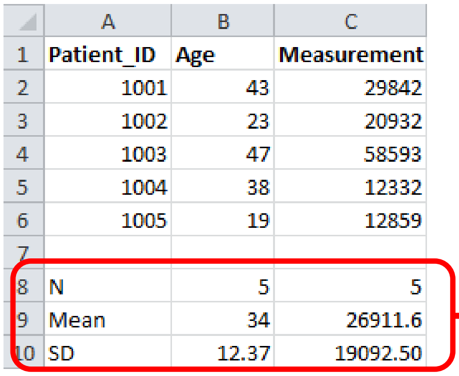


	A	B	C	D
1	**Exp_date**	**Mouse_ID**	**Treatment**	**Measurement**
2	6/10/2014	1001	Control	29842
3	6/10/2014	1002	Control	20932
4	6/10/2014	1003	Drug	58593
5	6/10/2014	1004	Drug	12332
6	6/12/2014	1005	Control	85669
7	6/12/2014	1006	Control	58954
8	6/12/2014	1007	Drug	15658
9	6/12/2014	1008	Drug	39987
10	6/12/2014	1009	Drug	69888
11	6/13/2014	1010	Control	56897
12	6/13/2014	1011	Drug	57986
13	6/13/2014	1012	Drug	89876

All three batches are now combined, with one column for each variable. Each mouse has an '`Exp_date`' value to identify group membership.

11. Avoiding Common Sheet Pitfalls

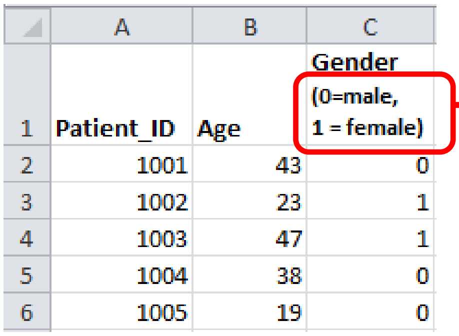
- Do not include any plots/figures on your data sheet – put them on a separate sheet.
- Do not include notes or summary statistics (e.g., means, standard deviations) next to or below your data on the same sheet!



	A	B	C
1	Patient_ID	Age	Measurement
2	1001	43	29842
3	1002	23	20932
4	1003	47	58593
5	1004	38	12332
6	1005	19	12859
7			
8	N	5	5
9	Mean	34	26911.6
10	SD	12.37	19092.50

NO

- Do not put variable descriptions or information about the values of a variable in the same cell as the variable name/header.

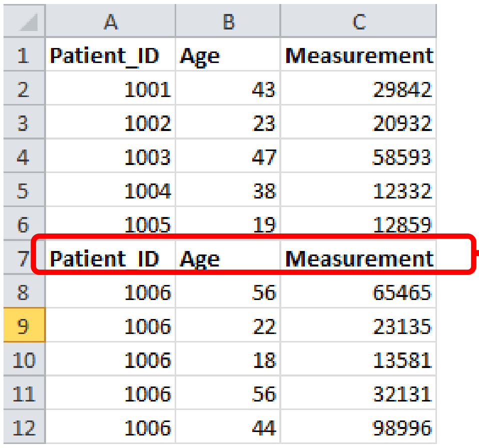


	A	B	C
1	Patient_ID	Age	Gender (0=male, 1 = female)
2	1001	43	0
3	1002	23	1
4	1003	47	1
5	1004	38	0
6	1005	19	0

NO.

Please see '1. Variable Names', page 2 and '3. General Data Entry Rules', page 4 for examples of how to include extra information about a variable.

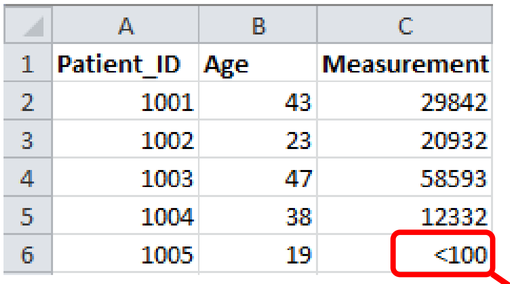
- Do not repeat headers throughout the worksheet.



	A	B	C
1	Patient_ID	Age	Measurement
2	1001	43	29842
3	1002	23	20932
4	1003	47	58593
5	1004	38	12332
6	1005	19	12859
7	Patient_ID	Age	Measurement
8	1006	56	65465
9	1006	22	23135
10	1006	18	13581
11	1006	56	32131
12	1006	44	98996

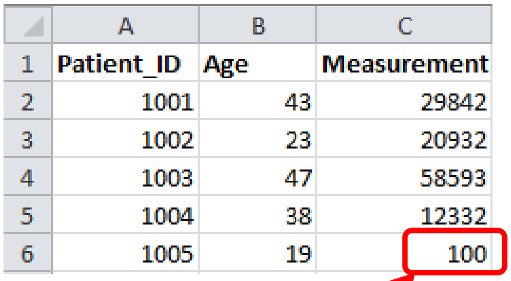
NO

- If the variable is numeric, do not use ‘<’ or ‘>’.



	A	B	C
1	**Patient_ID**	**Age**	**Measurement**
2	1001	43	29842
3	1002	23	20932
4	1003	47	58593
5	1004	38	12332
6	1005	19	<100

NO



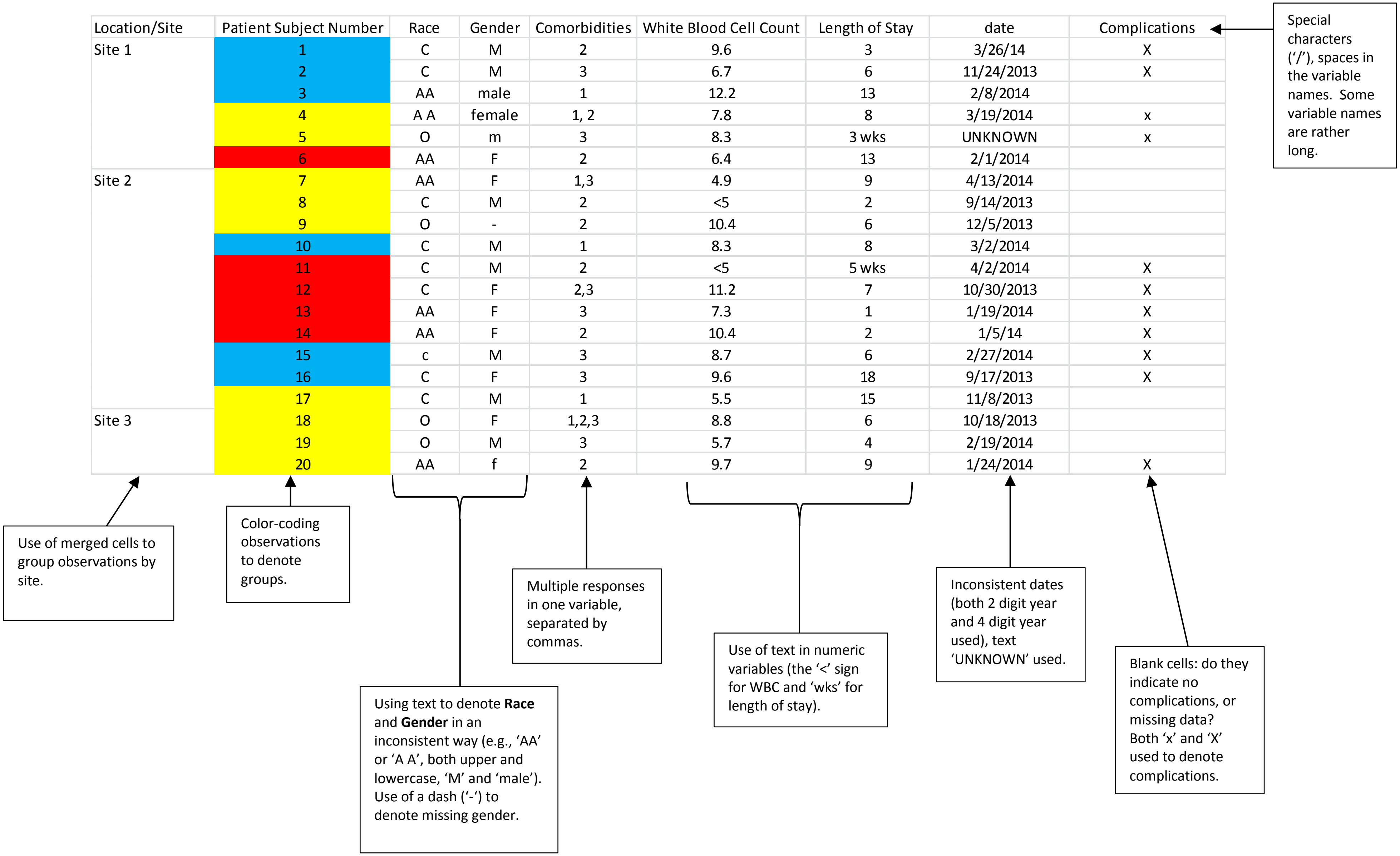
	A	B	C
1	**Patient_ID**	**Age**	**Measurement**
2	1001	43	29842
3	1002	23	20932
4	1003	47	58593
5	1004	38	12332
6	1005	19	100

YES

Instead of ‘<100’, replace with the lower bound of 100 in this example. **Talk with your statistician** about what makes clinical sense for your data!

Example: Spreadsheet with Some Common Data Issues

The following hypothetical spreadsheet would require extensive data management before analysis.



Location/Site	Patient Subject Number	Race	Gender	Comorbidities	White Blood Cell Count	Length of Stay	date	Complications
Site 1	🟦 1	C	M	2	9.6	3	3/26/14	X
	🟦 2	C	M	3	6.7	6	11/24/2013	X
	🟦 3	AA	male	1	12.2	13	2/8/2014	
	🟨 4	A A	female	1, 2	7.8	8	3/19/2014	x
	🟨 5	O	m	3	8.3	3 wks	UNKNOWN	x
	🟥 6	AA	F	2	6.4	13	2/1/2014	
Site 2	🟨 7	AA	F	1,3	4.9	9	4/13/2014	
	🟨 8	C	M	2	<5	2	9/14/2013	
	🟨 9	O	-	2	10.4	6	12/5/2013	
	🟦 10	C	M	1	8.3	8	3/2/2014	
	🟥 11	C	M	2	<5	5 wks	4/2/2014	X
	🟥 12	C	F	2,3	11.2	7	10/30/2013	X
	🟥 13	AA	F	3	7.3	1	1/19/2014	X
	🟥 14	AA	F	2	10.4	2	1/5/14	X
	🟦 15	c	M	3	8.7	6	2/27/2014	X
	🟦 16	C	F	3	9.6	18	9/17/2013	X
	🟨 17	C	M	1	5.5	15	11/8/2013	
Site 3	🟨 18	O	F	1,2,3	8.8	6	10/18/2013	
	🟨 19	O	M	3	5.7	4	2/19/2014	
	🟨 20	AA	f	2	9.7	9	1/24/2014	X

Special characters ('/'), spaces in the variable names. Some variable names are rather long.

Use of merged cells to group observations by site.

Color-coding observations to denote groups.

Using text to denote **Race** and **Gender** in an inconsistent way (e.g., 'AA' or 'A A', both upper and lowercase, 'M' and 'male'). Use of a dash ('-') to denote missing gender.

Multiple responses in one variable, separated by commas.

Use of text in numeric variables (the '<' sign for WBC and 'wks' for length of stay).

Inconsistent dates (both 2 digit year and 4 digit year used), text 'UNKNOWN' used.

Blank cells: do they indicate no complications, or missing data? Both 'x' and 'X' used to denote complications.

Example: Spreadsheet with Data Issues – FIXED!

The following hypothetical spreadsheet is now ready for analysis by a statistician.

site	patient_ID	group	race	gender	comorbid_1	comorbid_2	comorbid_3	WBC	LOS	date	complications
1	1	1	C	M	0	1	0	9.6	3	3/26/2014	1
1	2	1	C	M	0	0	1	6.7	6	11/24/2013	1
1	3	1	AA	M	1	0	0	12.2	13	2/8/2014	0
1	4	2	AA	F	1	1	0	7.8	8	3/19/2014	1
1	5	2	O	M	0	0	1	8.3	21		1
1	6	3	AA	F	0	1	0	6.4	13	2/1/2014	0
2	7	2	AA	F	1	0	1	4.9	9	4/13/2014	0
2	8	2	C	M	0	1	0	4.9	2	9/14/2013	0
2	9	2	O		0	1	0	10.4	6	12/5/2013	0
2	10	1	C	M	1	0	0	8.3	8	3/2/2014	0
2	11	3	C	M	0	1	0	4.9	35	4/2/2014	1
2	12	3	C	F	0	1	1	11.2	7	10/30/2013	1
2	13	3	AA	F	0	0	1	7.3	1	1/19/2014	1
2	14	3	AA	F	0	1	0	10.4	2	1/5/2014	1
2	15	1	C	M	0	0	1	8.7	6	2/27/2014	1
2	16	1	C	F	0	0	1	9.6	18	9/17/2013	1
2	17	2	C	M	1	0	0	5.5	15	11/8/2013	0
3	18	2	O	F	1	1	1	8.8	6	10/18/2013	0
3	19	2	O	M	0	0	1	5.7	4	2/19/2014	0
3	20	2	AA	F	0	1	0	9.7	9	1/24/2014	1

Short, concise variable names with no spaces or special characters. Use of underscore ('_') to separate words.

Instead of merged cells, each patient has site listed individually.

Replaced color-coding with a new variable to identify group.

Race and **Gender** entries are now consistent (for example, 'C' alone denotes Caucasian). A blank cell is used to indicate missing data.

Instead of multiple responses in one variable, created separate 0/1 variables for each possible response (0=does not have comorbidity, 1 = has comorbidity).

Removed all text from numeric variables. Replaced '<5' with '4.9' after discussions between investigator and statistician.

Dates are all now of the form MM/DD/YYYY. Replaced 'UNKNOWN' with blank cell.

Changed to 0/1 variable instead of text. Patients without complications are now denoted as '0' to avoid being confused with missing data.